(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 992 922 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
12.04.2000 Bulletin 2000/15

(51) Int Cl.[7]: **G06F 17/30**

(21) Application number: **99307580.3**

(22) Date of filing: **24.09.1999**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **29.01.1999 US 236622**
02.10.1998 US 102944 P

(71) Applicant: **International Business Machines Corporation**
Armonk, NY 10504 (US)

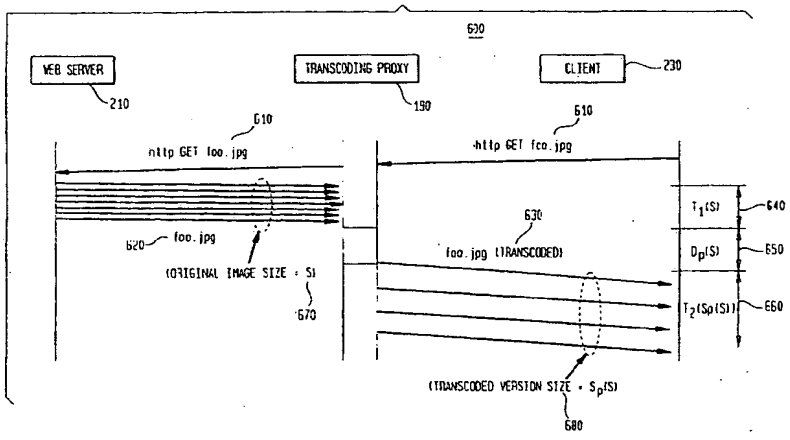
(72) Inventors:
- **Bhagwat, Pravin, c/o IBM United Kingdom Ltd.**
  Winchester, Hampshire SO21 2JN (GB)
- **Han, Richard Yeh-whein,c/o IBM United Kingdom Ltd.**
  Winchester, Hampshire SO21 2JN (GB)
- **La Maire, Richard O.,**
  c/o IBM United Kingdom Ltd.
  Winchester, Hampshire SO21 2JN (GB)
- **Mummert, Todd William,**
  c/o IBM United Kingdom Ltd.
  Winchester, Hampshire SO21 2JN (GB)
- **Rubas, James, c/o IBM United Kingdom Ltd.**
  Winchester, Hampshire SO21 2JN (GB)

(74) Representative: **Davies, Simon Robert**
IBM,
United Kingdom Limited,
Intellectual Property Law,
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(54) Automatic image data quality adjustment to reduce response time of a Web server

(57) The present invention provides methods, devices and systems for dynamically adjusting transcoding parameters so as to increase the benefits of transcoding. Methods of adaptation are designed to cope with the variability of network characteristics and of the size of transcoded images. The invention also provides a method and apparatus to enable the transcoding proxy to adjust a quality-size tradeoff on a per-image and/or a per-client basis. The adaptive transcoder chooses different parameters for each object, and provides performance improvements. The invention further provides a general framework for making policy decisions taking into account available bandwidth, content and type of image, and user preferences. The invention also includes methods for generating feedback about the choice of optimal transcoding parameters to the user.

FIG. 6



Description

[0001] This invention relates to browser proxies that transform image objects embedded inside documents. It is more generally related to any proxy or gateway system that processes MIME encapsulated objects and the world wide web.

[0002] Transcoding proxies sometimes act as intermediaries between web servers and a variety of client devices that are connected over communication links with widely varying characteristics. Generally, a transcoding proxy provides two major benefits: dramatic reduction in download times for low-bandwidth links, and tailoring of web data to the client device.

[0003] Mobile devices are frequently connected via low-bandwidth wireless or medium bandwidth wireline modem links that make the viewing of rich web content, such as images, very cumbersome due to the very long download times that result. In addition, the cost of such downloads can be prohibitive over tariffed wide-area wireless networks. Transcoding web proxies can reduce the size of web data while maintaining most of its semantic value. Download time reductions of six to ten times may be achieved for many typical web images without losing their intelligibility.

[0004] The second important benefit is that such an intermediate proxy is sometimes also capable of tailoring text and images for the multitude of small, weakly connected, but web-enabled, mobile devices that are now available. The capabilities of these mobile devices to receive, process, store and display web content varies widely. Given the variety of client devices, it is difficult for Internet content publishers to tailor content for the individual devices. Consequently, an active web proxy is used to transcode/change the web content to best fit the resolution, color-depth, and size constraints of a small-screened device and to reduce the size of the stored data to a fraction of its original size.

[0005] A typical transcoding web proxy may be implemented by adding a transcoding module in an HTTP proxy. The transcoding module is an input/output system that takes a data object (e.g., an image or HTML page) as input and transforms it into a low resolution object (e.g., a low-quality image or summarized HTML page) according to a set of chosen parameters. These parameters are input to the transcoder, which determines the resolution loss and size reduction of the object. The following discussion is mostly concerned with image objects.

[0006] To maximize the benefits of transcoding, it is important to choose the quality Vs. size tradeoff such that the best quality image is sent within the delay tolerance specified by the user. Choosing the right set of transcoding parameters, however, is difficult. First, it is hard to estimate the size of the transcoded image size because the degree of compression achieved through transcoding is content dependent. For example, some JPEG files can be compressed by up to 80% while others may only yield less than 5% compression. The same is true when GIF-to-GIF or GIF-to-JPEG transformations are applied. Since transcoding is a compute-intensive process, processing is only worthwhile when high compression ratios can be achieved. For images that are already well-compressed (or of low quality) transcoding may be wasteful.

[0007] Even if the size of the transcoded output can be predicted, network variability makes it hard to predict image download times. Since the user perceived quality and latency are the final performance measures, accurate estimates of download times are needed in order to choose the optimal point in the quality Vs. size tradeoff. Over best effort service networks, such as the Internet, accurate estimates of network characteristics (bandwidth, delay, loss rate) are hard to compute. Often the variance is so large that the statistical estimates are not meaningful for making any quality adjustments.

[0008] Another issue that makes policy decisions a complicated task is the location of the bottleneck in the network. when the path between the web server and the proxy is the bottleneck, transcoding does not help at all. On the contrary, it makes things worse due to the store and forward nature of the operation. Ideally, web proxies should stream images from web servers to clients whenever the server to proxy path is the bottleneck. Even in cases where the proxy to client path is the bottleneck, any benefit due to size reduction must be weighed against the store and forward delay incurred due to transcoding.

[0009] Due to the difficulty of estimating the network characteristics and image content, transcoding proxies may often only support adaptation at a very coarse level. Adaptation involves changing transcoding parameters which are usually selected from a predefined set of defaults, or chosen by the user. The problem with both approaches is that the full potential of transcoding is not utilized. Predefined static policies are unresponsive to changes in network bandwidth and user selectable policy decisions are likely to be sub-optimal due to the lack of knowledge about the state of the network. No method exists for providing feedback to the user about the combined state of the network and the proxy. Using iterative refinement, users can converge to the optimal choice of transcoding parameters, but such a process is time consuming and not very user-friendly.

[0010] Figure 1 shows an example transcoding proxy scenario 100 with varied links and varied client devices. As shown in Figure 1, we consider an example scenario in which all of the web requests and responses for clients 130-134 are passed through a generic (Hypertext Transfer Protocol) HTTP proxy 190. In the example scenario of Figure 1, it is the downloading of large data objects, primarily images, over the last link 160-164 from the proxy to the client that typically is the main source of delay in the end-to-end response time experienced by the clients 130-134.

[0011] The transcoding proxy 190 provides massive data thinning of web data to enable real-time browsing of web

data over low-speed wide-area wireless links, like Cellular Digital Packet Data (CDPD) 163 that provides 10 kb/s or less of throughput, depending upon the number of clients sharing the link. The end client could be a full-function PC 134 with a good color display, so that the primary problem is bandwidth reduction. Alternatively, the client might be a small, web-enabled, mobile devices 130-132 in which case it is advantageous to tailor the web data for the specific client device, particularly the client's display characteristics. Different aspects of transcoding proxy design benefit different scenarios (i.e., tariffed proxy-client link, strong/weak client display, etc.), but all are handled well within the same proxy architecture 100.

[0012] Figure 2 shows a block diagram of an embodiment of transcoding proxy 190 used to transform objects based on user specified preferences and static policies. A transcoding proxy 190 is built by combining a transcoding module 240 with an HTTP proxy engine 220. An HTTP request 222 originates from a client 230 and is forwarded 224 by the proxy 190 to a web server 210. The response data 226 (i.e., HTML pages and GIF and JPEG images) are transformed by the transcoder 240 and then forwarded 228 to the client 230. Typically, a number of transcoding parameters are specified to the transcoder 240 in order to achieve the desired quality/size reduction of the object contained in the response data 226. Transcoding proxies in use today either use a static set of policies 250 or use some form of user specified preferences 260 via path 265 to determine the transcoding parameters. When a fixed set of transcoding parameters are applied to all objects, results are not always beneficial. In fact, in many cases, transcoding leads to poorer performance.

[0013] Accordingly, in a first aspect, the present invention provides a method for a transcoding proxy to facilitate browsing between a plurality of client devices and a plurality of servers connected via a communication network, the method comprising: receiving a request from one of the client devices for an object stored at one of the servers, forwarding the request for the object to said one of the servers, receiving the object from said one of the servers, examining preferences specified by a user of said one of the client devices, examining contents of the object, examining communication network characteristics, choosing a set of transcoding parameters, forming a transcoded form of the object, and sending the transcoded form to said one of the clients.

[0014] Preferably, the network characteristics include bandwidth, and examining network characteristics includes estimating network bandwidth between said one of the servers and the proxy as well as between the proxy and said one of the clients.

[0015] Preferably, the network characteristics include delay, and the step of examining network characteristics includes estimating delay between said one of the servers and the proxy as well as delay between the proxy and said one of the clients.

[0016] The method preferably further comprises providing feedback to the user about a level of transcoding performed on the object to form the transcoded form.

[0017] Preferably, the step of examining includes determining the size of the object.

[0018] In the method as described the object may be of type image forming an image object, and the method further comprises: determining dimensions of the image object, and calculating the compression ratio of the image object. Further to be preferred is that the dimensions of the image object are determined by area of the image in square pixels, and the compression ratio is determined by the bpp ratio of the image object.

[0019] The step of forming a transcoded form preferably employs dynamic adaptation.

[0020] The step of forming a transcoded form is preferably started before the step of receiving the object from said one of the servers is complete. The received object type is preferably of type JPEG forming a JPEG object. Further to be preferred is that the step of forming a transcoded form includes performing JPEG-to-JPEG image transcoding and the step of sending the transcoded form starts writing out at least one MCU of JPEG-encoded output image data before the step of receiving the object is complete.

[0021] In the method as described, the step of sending out the transcoded form is preferably started after processing an initial fraction of the received object, and before the step of receiving the image object from said one of the servers is complete.

[0022] The step of sending out the transcoded form is preferably started before the step of forming a transcoded form of the object is complete.

[0023] In a second aspect, the present invention provides a method for a proxy to form a transcoded form of an object received from a server in satisfaction of a request from a client for an object available from the server, the method comprising dynamically adapting parameters for transcoding the object for the client, forming a transcoded form of the object, and sending the transcoded form to the client. The step of adapting parameters preferably includes determining at least one characteristic of the object. One characteristic is preferably an object-header, the object-header providing information about the size and the type of the object. The method preferably further comprises comparing the size of the object to a threshold parameter called "size_threshold". The step of adapting parameters preferably includes gathering present network characteristics between the server and the proxy and between the proxy and the client. One of the characteristics is preferably network bandwidth and the step of adapting includes estimating network bandwidth between the server and the proxy as well as between the proxy and the client.

[0024] In the method as described, the transcoded form is preferably dependent upon the estimated bandwidth. The step of adapting also preferably includes retrieving preferences of the user, and wherein the transcoded form is dependent upon the preferences. The step of adapting preferably includes examining the contents of the object. The object is preferably of type image forming an image object, and examining the content of the image object includes determining dimensions of the image.

[0025] In the method as described, the step of adapting is preferably dependent upon determining the compression ratio of the image object. The type of image object is further preferably GIF and wherein the step of adapting is dependent upon comparing the compression ratio against a predetermined policy threshold called "gif_threshold".

[0026] The method as described preferably further comprises predicting at least one parameter of the transcoded form of the object. The transcoded form is further preferably the same as an original form of the object. Preferably, at least one of the parameters includes a size of the transcoded form. Preferably also, at least one of the parameters includes the time spent in transcoding the object.

[0027] In a third aspect, the present invention provides a method for predicting parameters of a transcoded form of an object, the object having an initial size and dimension, and the object being received from a server in satisfaction of a request from a client for the object, the method comprising: computing the bpp ratio of the object, gathering a set of statistics of a plurality of previously transcoded objects, and employing the set of statistics and the bpp ratio for predicting the parameters.

[0028] In a method as described, at least one of the parameters is preferably size and the set of statistics includes sizes of a plurality of previously transcoded objects statistics.

[0029] The object is preferably of type image and the set of statistics includes image quality. The plurality of previously transcoded objects are preferably chosen from a predetermined benchmark suite of images.

[0030] In a method as described, the step of employing preferably uses dynamically updating the set of statistics using the statistics of the currently transcoded object. At least one of the parameters is preferably a duration for forming a transcoded form of an object and the set of statistics includes the duration for forming a plurality of previously transcoded objects.

[0031] In a fourth aspect, the present invention provides a transcoding proxy system for facilitating browsing between a plurality of clients and a plurality of servers connected via a communication network, the proxy comprising an HTTP proxy engine to receive a request from one of the clients for an object stored at one of the servers, and to fetch the object from said one of the servers, an object transcoder to form a transcoded form of the object using a set of parameters for transcoding, a dynamic policy module to determine the set of parameters of transcoding, an image size and delay predictor module to gather characteristics of the object, a user preference module to gather quality preferences specified by a user of said one of the clients, and a bandwidth estimation module to estimate available network bandwidth, wherein dynamic policy module dynamically adjusting the parameters of transcoding using the input received from the image size and delay predictor module, user preferences module, and bandwidth estimation module for the purpose of improving satisfaction for the user, and the transcoding system providing feedback to the user about the level of transcoding performed.

[0032] The user preferences module preferably further collects the characteristics such as display size, resolution, & CPU speed of said one of the devices, and provides those characteristics to the dynamic policy module.

[0033] In the system as described, the bandwidth estimation module preferably collects traces of previously established network connections between said one of the servers & the proxy, collects traces of previously established network connections between the proxy and said one of the clients, and estimates the object download time by performing statistical analysis on the collected traces.

[0034] The statistical analysis used for estimating bandwidth between said one of the servers and the proxy is preferably based on computing a statistical measure such as median, mean, or mode of download times of previously fetched objects as determined from the collected traces.

[0035] The statistical analysis used for estimating bandwidth between the proxy and said one of the clients is preferably based on computing aggregate bandwidth of all active connections between the proxy and said one of the clients. The system preferably further comprises displaying a slide bar on said one of the client's display for collecting the user specified preferences. In a system as described, the user of said one of the clients can preferably specify the tradeoff between download time and data quality through the use of a graphical user interface with a slide bar. The user of said one of the clients can preferably specify through the use of a graphical user interface with a slide bar the tradeoff between download time and image quality including a specific switch to select color or gray scale as the desired output form. The user of said one of the clients can further preferably specify through the use of a graphical user interface with a slide bar, the desire to maintain a target response time such that the system automatically reduces data quality (and hence data download size) to compensate for dynamic variations in bandwidth to said one of the clients. The graphical user interface slider bar is also preferably used as an output interface for showing the optimal choice of transcoding parameters to the user.

[0036] In a fifth aspect, the present invention provides an article of manufacture comprising a computer usable me-

dium having computer readable program code means embodied therein for causing dynamic adaptation of transcoded form of an object in a transcoding proxy, the computer readable program code means in said article of manufacture comprising computer readable program code means for causing a computer to effect: a proxy receiving an object associated with a user from a server, determining parameters of the object, retrieving preferences of the user, gathering present network characteristics, obtaining transcoding policy thresholds, performing a policy decision based upon object parameters, user preferences, network characteristics, and policy thresholds, forming a transcoded object, providing feedback of a level of transcoding performed on the object to the user, and sending the transcoded object to the user.

[0037] In a sixth aspect, the present invention provides an article of manufacture comprising a computer usable medium having computer readable program code means embodied therein for causing a transcoding proxy to facilitate browsing between a plurality of client devices and a plurality of servers connected via a communication network, the computer readable program code means in said article of manufacture comprising computer readable program code means for causing a computer to effect: receiving a request from one of the client devices for an object stored at one of the servers, forwarding the request for the object to said one of the servers, receiving the object from said one of the servers, examining preferences specified by a user of said one of the client devices, examining contents of the object, examining communication network characteristics, choosing a set of transcoding parameters, forming a transcoded form of the object, and sending the transcoded form to said one of the clients.

[0038] The computer readable program code means in said article of manufacture preferably further comprises computer readable program code means for causing a computer to effect providing feedback to the user about a level of transcoding performed on the object to form the transcoded form. The computer readable program code means in said article of manufacture preferably further comprises computer readable program code means for causing a computer to effect determining dimensions of the object, and calculating the compression ratio of the object. The computer readable program code means in said article of manufacture preferably further comprises computer readable program code means for causing a computer to effect starting the step of forming a transcoded form before the step of receiving the object from said one of the servers is complete.

[0039] In a seventh aspect, the present invention provides a computer program product comprising a computer usable medium having computer readable program code means embodied therein for causing a proxy to form a transcoded form of an object received from a server in satisfaction of a request from a client for an object available from the server, the computer readable program code means in said computer program product comprising computer readable program code means for causing a computer to effect: dynamically adapting parameters for transcoding the object for the client, forming a transcoded form of the object, and sending the transcoded form to the client.

[0040] The computer readable program code means in said computer program product preferably further comprises computer readable program code means for causing a computer to effect gathering present network characteristics between the server and the proxy and between the proxy and the client. The computer readable program code means in said computer program product preferably further comprises computer readable program code means for causing a computer to effect adapting parameters for transcoding based upon the estimated bandwidth and preferences of the user.

[0041] In an eighth aspect, the present invention provides a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for predicting parameters of a transcoded form of an object, the object having an initial size and dimension, and the object being received from a server in satisfaction of a request from a client for the object, said method steps comprising: computing the bpp ratio of the object, gathering a set of statistics of a plurality of previously transcoded objects, employing the set of statistics and the bpp ratio for predicting the parameters.

[0042] Said method step of employing the set of statistics preferably further comprises updating the set of statistics using the statistics of the currently transcoded objects.

[0043] An aspect of the present invention is to provide methods, devices and systems for dynamically adjusting transcoding parameters so as to increase the benefits of transcoding. Methods of adaptation are designed to cope with the variability of network characteristics and of the size of transcoded images.

[0044] In an embodiment, the invention includes three new components: an image size predictor, a network bandwidth (b/w) analyzer, and a policy module. Before initiating any transcoding action, the policy module queries the image size predictor to estimate the size of the output image. The b/w analyzer is queried to collect an estimate of the image transmission time from the server to the proxy, and from the proxy to the client. Based on the collected estimates, the proxy decides whether or not to transcode the image. In addition, the policy module can also compute the optimal point in the quality vs. size tradeoff that would provide the user specified performance criteria (e.g., reduced response time, increased quality).

[0045] Another aspect of the present invention is to provide a method and/or apparatus to enable the transcoding proxy to adjust a quality-size tradeoff on a per-image and/or a per-client basis. The adaptive transcoder chooses different parameters for each object, and provides performance improvements.

[0046] Still another aspect of this invention provides a general framework for making policy decisions taking into account available bandwidth, content and type of image, and user preferences. The administrator of the proxy chooses from a variety of optimization objectives so as to obtain improved performance from the system. In one embodiment when the proxy transcoder throughput is the bottleneck, the policy module is instructed to judiciously use CPU resources so as to reduce the response time for all users. An advantageous element of the invention is the automated nature of decision making, to free up users from actively controlling the policy engine of the proxy.

[0047] In still another aspect of the present invention a method is provided for generating feedback about the choice of optimal transcoding parameters to the user. In an example of embodiment, the transcoding system provides feedback to the user by dynamically adjusting the position of the user preference slider bar. The user preference slider bar serves both as an input as well as an output device.

[0048] A preferred embodiment of the present invention will now be described by way of example, with reference to the drawings in which:

FIG. 1 shows an example transcoding proxy scenario with varied links and varied client devices;

FIG. 2 shows a block diagram of a transcoding proxy used to transform objects based on user specified preferences and static policies;

FIG. 3 shows a block diagram of an example of a transcoding proxy modified in accordance with the present invention to include an image size predictor, bandwidth estimator, dynamic policy module, and a user feedback generator;

FIG. 4 shows an example block diagram of an example HTTP proxy with caching and transcoding modules in accordance with the present invention;

FIG. 5 shows an example flow diagram of a transcoding dynamic policy module in accordance with the present invention;

FIG. 6 shows an example web request-response cycle using a transcoding proxy in accordance with the present invention;

FIG. 7 shows an example regime where transcoding is useful in accordance with the present invention;

FIG. 8 shows an example flow diagram of a example policy function in accordance with the present invention;

FIG. 9 shows an example block diagram of an example image size prediction module in accordance with the present invention;

FIG. 10 shows an example block diagram of an example bandwidth prediction module in accordance with the present invention;

FIG. 11 shows input/feedback user interface in accordance with the present invention; and

FIG. 12 illustrates an example timing diagram of streaming transcoding in accordance with the present invention.

[0049] Figure 3 shows a block diagram 300 of an example embodiment of changes to transcoding proxy 190 in accordance with the present invention. A comparison of Figure 3 to Figure 2 shows a replacement of the static policy module, 250 in Figure 2, with a dynamic policy module, 370 in Figure 3. A purpose of the dynamic policy module 370 is to make decisions concerning when to turn transcoding on and off and what transcoding policy (i.e., the transcoding algorithm along with its parameters) to use. The dynamic policy module 370 also interfaces with an image size and delay predictor 375, a bandwidth estimator 380, and a user feedback provider 390. In the embodiment shown, the policy module 370 employs a number of criteria, including: the characteristics of the data (e.g., size of the images, current encoding efficiency, structural role in the HTML page) as determined by the content analysis flow diagram (shown in Figure 5), the current estimate of the bandwidth on the proxy-to-client and server-to-proxy links (shown in Figure 10), the characteristics of the client, particularly the client display capabilities, and the user preferences concerning the preferred rendering of the data (shown as the user slide bar preferences in Figure 11).

[0050] The items shown in Figures 5, 10 and 11 are described below. In particular, the user *slide bar preferences* of Figure 11, provide a method of interacting with the transcoding proxy so as to dynamically change the tradeoff between

image quality and download time. In addition to serving as an input interface, the slider bar (1140 1150 1160 of Figure 11) also acts as an output interface, displaying the feedback 390 that is received from the dynamic policy module 370.

[0051] Figure 4 shows a block diagram 400 illustrating an example function of a multi-resolution cache 410 in a transcoding HTTP proxy system 400 in accordance with the present invention. Caches are useful in HTTP proxies to provide reduced response time for repeated data requests (by the same or different clients) for the same data object. A variety of methods may be used to assure that the cached data is up-to-date. In the example caching transcoding system 400, the multi-resolution cache 410 stores an original version of the data, as well as other possible forms of the data, including reduced resolution versions that have been transcoded for specific device types.

[0052] As an example, we again consider the case of an image data object, but it is understood by those skilled in the art that these methods can be applied to other data types. We describe a method for storing, tagging and retrieving various forms of the data in the context of a caching transcoding proxy system.

[0053] Referring to Figure 4, it is assumed that a JPEG image is received from the server 210 in response to a request by a client 230. This image is decoded from the lossy JPEG encoding format to a bit-map representation of the image. Those skilled in the art of image processing recognize that the JPEG encoding standard describes images using coefficients that weight Discrete Cosine Transforms (DCTs) so a compute-intensive decoding step is generally required to obtain the actual colors of each pixel. This bit-map requires a larger storage size than the original JPEG image. However, given an adequate size cache it may be worthwhile to store (via data path 420) this expanded size image so as to avoid the compute-intensive step of JPEG decoding when a transcoded version of the same image is later required but with different transcoding parameters than the first request. In a second step, based on the transcoding parameters (i.e., color depth, scaling factor, and the JPEG quality factor) the JPEG image is re-encoded. The image may be re-encoded as a JPEG image or in an alternative encoding format. The final transcoded version of the image is also stored in the multi-resolution cache 410 via data path 420.

[0054] In the example embodiment, when additional data requests occur, the HTTP proxy 220 first checks its cache 410 to see if an "up-to-date" version of the data object is available at the requested resolution or transcoding level. Each object in the multi-resolution cache 410 is stored with a version specifier that includes: a URL description, a time stamp for the data object, and the object characteristics. For a JPEG image the object characteristics include color-depth, scaling factor, and JPEG quality factor. An alternative embodiment has an indication that the JPEG image has been decoded and stored in its bit-map form, or that it has been converted and stored as a GIF with various characteristics.

[0055] If an up-to-date version of the object is available with the requested type and resolution, then that version is returned to the client. If this is not available, but an up-to-date version of the object exists in either its original JPEG form or the decoded bit-map form, then this version is returned to the transcoder 240 with an indication of its characteristics. This enables the transcoder 240 to produce the desired version of the object, which is returned to the client 230 and stored in the multi-resolution cache 410 via data path 420.

[0056] It is noted that there are several extensions to this scheme that are obvious to those skilled in the art. One extensions uses an already transcoded version of the object, rather than to the original data object, to generate a further resolution-reduced version of the object. Methods for managing the different resolution versions of data objects in a cache are further described in R. O. LaMaire and J. T. Robinson, "Conserving Storage Space by Means of Low Resolution Objects", docket Y0997308, US patent application filed February 13, 1998.

[0057] Figure 5 shows an example flow diagram for the dynamic policy module 370 of Figures 3 and 4. Figure 5 also shows how the policy module 370 interfaces with the HTTP proxy engine 220 and the object transcoder 240 of Figure 3. The methods described below apply to many content types including text, images, audio, and video. However, the following discussion focuses on image data types only. It is evident to those skilled in the art that concepts and dynamic policies are applicable to other media types.

[0058] Figure 5 shows that based on the response header received from the server 210, the HTTP proxy engine 220 first determines the size of the object 510. If the content-type of the response is "image/*" 520, the proxy engine 220 passes the handle for the object to the dynamic policy module 370 for further analysis. Inside the policy module 370, the size of the input object is compared against a pre configured threshold called "size_threshold" 530. If the objects is smaller than "size_threshold", or the content type is not "image/*" 520, then the object is not transcoded, but instead is forwarded 515 to the client without any content modification. Small objects (such as bullets, thumbnails, logos, etc.) found on the web are typically GIF objects which are already well compressed due to GIF encoding. Transcoding such objects does not generally yield further compression.

[0059] We convert GIF images to GIF or JPEG images that are reduced in size and/or color-depth (the choice of GIF or JPEG as the end format depends on the image characteristics). In addition, we convert JPEG images to JPEG images that are reduced in JPEG quality, size, and/or color-depth. JPEG quality refers to a transcoding parameter that is used to determine the degree to which the coefficients of the Discrete Cosine Transformations used in the JPEG encoding standard are quantized. It has been found that the JPEG quality parameter is also a good predictor of perceived image quality. This parameter varies in the range 1 to 100, where 100 represents very high quality. JPEG images

found on the web typically have a JPEG quality parameter of 75.

[0060] For large images, the type of the image coding (JPEG or GIF) and the efficiency of coding are important factors in transcoding decision making. Since JPEG is a lossy compression method, size reduction is always possible by reducing quality factor. Similar quality reduction, however, cannot be applied to GIF files since GIF is a lossless compression method. To achieve quality reduction, a GIF file must first be decoded and then re-encoded as a quality-reduced JPEG image. This method, however, may not always provide size reduction. GIF format is usually more efficient for coding maps, logos, and drawings while JPEG is more efficient for coding natural images. Converting GIF to JPEG is useful only when the original GIF image is not efficiently coded.

[0061] we define bits per pixel (bpp) as a measure of the compression efficiency. Bpp is computed as the ratio of the image file size to the image area in pixels. In processing step 540, X and Y dimensions and the bpp value of the input image are computed by parsing the image header. If the input object is of type "image/jpg" 550, transcoding is always performed. -However, if the content type is "image/gif" 525, only those objects which yield a bpp ratio larger than "gif_threshold" 535 are transcoded. GIF files that are not very efficiently encoded yield a bpp value that is larger than "gif_threshold". Thus, the decision step 535 is very effective in identifying compressible GIF files with high accuracy. Though not show in Figure 5, it is evident to those skilled in the art that other transcoding policies, such as scaling and file truncation (for progressively encoded data) can be used for well-compressed GIFs.

[0062] An important aspect of the proposed invention is that decision steps 510-565 are carried out as soon as the image header is received. If the decision is to not transcode, image segments can be forwarded as soon as they are received from the server without incurring store and forward delay. Similarly, when transcoding is to be performed, images can either be buffered and then transcoded (store and forward transcoding), or each segment can be transcoded on-the-fly (streaming transcoding method).

[0063] After identifying an image that is compressible, the next step involves determining the extent to which the selected image should be transcoded. The policy function 565 is responsible for collecting input from three different sources (image size predictor 375, bandwidth estimator 380, user preference selector 260) and subsequently selecting transcoding parameters in accordance with the steps shown in Figure 8. The chosen parameters determine the extent and types of compression performed by the object transcoder 240. For example, the scaling parameter determines how much an image is downsampled. Quantization parameters control how an image is quantized in the pixel domain and/or the frequency domain. The number of colors in a color mapped image can be reduced, or a 24-bit color image may be converted to 8-bit grayscale, or even a monochrome representation. The process of transcoding is performed in step 570 and the output of the transcoder is forwarded to the client 230.

[0064] An important aspect of the policy function 565 is the analytical framework for making transcoding decisions. The analytical framework takes into consideration factors such as available bandwidth, type and size of the image, user preferences and provides an objective criteria for making transcoding decisions. As an example, we consider the objective of minimizing response time for the user, but it is understood by those skilled in the art that using the same framework other optimization criteria can also be applied. We describe a method for determining when it is beneficial to transcode, and to what extent transcoding should be applied. The embodiment described herein is referred to as dynamic adaptation of transcoding parameters.

[0065] Figure 6 shows an example web request-response cycle and the response time of fetching an object of size S through a store-and-forward transcoding proxy 190. We define a store-and-forward image transcoder as an image transcoder which must wait to accumulate an entire input image before transcoding can begin on this image and then must wait to generate a transcoded image in its entirety before it is made available to be output. As shown in Figure 6, the original image 620 of size S (bytes) 670 is downloaded into the store-and-forward proxy over the server-proxy connection with effective bandwidth B_{sp} (bits/sec). The transcoder introduces a delay $D_x(S)$ 650 and generates an output image 630 of size $S_x(S)$ 680. Both the transcoding delay 650 and output image's byte size 680 are denoted to be dependent upon the input image's byte size S 670. The transcoded image is then transmitted over a proxy-client connection having effective bandwidth B_{pc}.

[0066] The policy function needs to weigh the cost (delay) of transcoding against any size reduction achieved by transcoding. For transcoding to provide benefits, delay introduced due to transcoding must be offset by the reduction in transmission time due to compression. For very low bandwidth proxy-client access links, the reduction in response time due to aggressive image compression typically far outweighs the addition to response time caused by compute-intensive transcoding. However, Figure 7 shows that as the bandwidth of the proxy-client link increases, there comes a point (transcoding threshold 710) at which it is no longer beneficial to transcode since the reduction in response time due to aggressive compression decreases as a function of the bottleneck link's bandwidth, while the transcoding time remains constant.

[0067] Suppose R_o is the response time of fetching a web object of size S from the web server with transcoding turned off. Similarly, let R_p denote the response time of fetching the transcoded version of the same web object through the transcoding proxy. For the purpose of the following discussion we assume that caching is not supported at the proxy.

[0068] The client perceived response time with transcoding turned off is the sum of the following three terms:

$$R_o = 2 * RTT_{pc} + 2 * RTT_{sp} + S/min(B_{pc}, B_{sp})$$

RTT_{pc} is the network roundtrip time latency between the client and the proxy and, similarly, RTT_{sp} is the between the proxy and the server. Fetching the web object requires a *TCP SYN/ACK* exchange as well as an equest/response, thereby contributing $2 * RTT_{pc} + 2 * RTT_{sp}$ to the delay term. In addition, a web image incurs nission delay equal to the spread in time between arrival of its first and the last bits. Let $min (B_{pc}, B_{sp})$ denote leneck bandwidth between the client and the server. In the absence of a proxy, the first and the last bits of an will be spread in time by $S/min (B_{pc}, B_{sp})$. This spread corresponds to the effective transmission time of the ver the concatenated server-to-proxy-to-client connection.

when transcoding is turned on, the proxy operates in a store and forward mode. $2 * RTT_{pc} + 2 * RTT_{sp}$ is ne fixed component of the response time. $D_p(S)$ is the additional term that represents the transcoding delay. ulting response time for the transcoded object can be expressed as:

$$R_p = 2 * RTT_{pc} + 2 * RTT_{sp} + D_p(S) + S/B_{sp} + Sp(S)/B_{pc}$$

oding will reduce response time if $R_p < R_o$. That is,

$$D_p(S) + S/B_{sp} + S_p(S)/B_{pc} < S/min(B_{pc}, B_{sp})$$

when $B_{pc} > B_{sp}$, R_p is always greater than R_o. On the other hand, when $B_{pc} < B_{sp}$, transcoding is useful if and

$$D_p(S) + S/B_{sp} < \{S - S_p(S)\}/B_{pc}$$

The above equation precisely characterizes the regime in which transcoding reduces response time. Figure s a flow diagram of an example policy function constructed using the analytical framework described above. rks the entry point of the policy function. The policy function 800 is called from 565 with the original image object one of the inputs to the policy function. Variable S is set equal to the input object size and the quality factor q ized to the best possible initial image quality (for example, the quality of the input image). In step 830, the policy 800 issues a query to the bandwidth estimator 380 asking for the estimated download time of the object size the specified server to the proxy, referred to as $T_1(S)$. It also asks for the estimate of the download time for the bject from the proxy to the client, referred to as $T_2(S)$. Based on the logs of previous connections to the chosen tion, the bandwidth estimator returns an estimate of $T_1(S)$ and $T_2(S)$. In the next step 840, the policy function the image size & delay predictor 375 to find an estimate of the transcoded image size $S_p(S)$. It then computes mated download time savings by subtracting the download time estimate of the transcoded image size $T_2(S_p$ m $T_2(S)$ 850. Finally, in step 870 the two quantities (transcoding delay + $T_1(S)$ - the download savings) and the esponse time reduction are compared. If the first term is less than the second term, computation is stopped chosen quality factor q is returned as an output 880. Otherwise, the loop 840-870 is reentered with a reduced

It is noted that more efficient search techniques or variants of objective functions can be designed by those in the art without departing from the spirit of the policy function framework presented in this invention. One on would be to rearrange the terms in the policy equation as follows:

$$\text{Response time reduction}(q) = \{S - S_p(S)\}/B_{pc} - D_p(S) - S/B_{sp}$$

In the above equation, $S_p(S)$ is also a function of the quality factor (the smaller the output size, the poorer the . Note that the transcoding delay has been found to be effectively independent of the quality factor. There are ependent variables in the above equation: q, the quality factor, and the target response time reduction. Several it policies can be developed within the framework of the above equation. For example:

nimize response time for all users;
ximize quality for a user specified response time constraint;
timize overall system performance, not just one user.

ə c = group image compression ratio G/G_p, which we assume to be on average equivalent to the overall image ression ratio. In summary, the streamed image transcoder should only perform transcoding when both Condition l Condition B are satisfied. If the proxy-server link is the bottleneck, i.e. $B_{cp} < B_{pc}$, then Condition B reduces to c where N is a number less than 1. Normally, the compression ratio is always greater than 1, so Condition B will 's be satisfied. In this case, only Condition A must be satisfied in order for transcoding to not be disadvantageous. t, when the proxy-server link is the bottleneck, Condition B could be interpreted as providing an upper bound on ıtio of expansion allowed for a transcoded image, namely $1/c < B_{pc}/B_{sp}$. Expansion of an image may occasionally cessary when format conversion is mandatory, e.g. GIF->Palm format. The above equation allows us to determine such format conversion will increase the chances of buffer overflow, and when format conversion will not cause overflow. For example, if $B_{sp} = 1\ b_{ps}$, $B_{pc} = 2\ b_{ps}$, and G = 1 bit, then Condition B says that the output group G_p xpand to a maximum of 2 bits. If the client-proxy link is the bottleneck, i.e. $B_{sp} > B_{pc}$, then Condition B says that ıage compression ratio c must be greater than the ratio of proxy-server to client-proxy bandwidths in order for coding to be worthwhile. In addition, Condition A must still be satisfied.

] It is noted that condition A and condition B are tight bounds that assume that the buffer must never be allowed erflow. Those skilled in the art recognize that looser constraints may be derived given that images are of finite-ı, rather than the continuous stream assumed in the analysis. More relaxed constraints would permit more time ınscoding and/or allow less aggressive compression.

] It is thus an aspect of the present invention to provide a method for a transcoding proxy to facilitate browsing ən client devices and servers connected via a communication network. The method includes receiving an HTTP equest from a client devices for an object stored at one of the servers, forwarding the GET request for the object server, receiving the object from the servers, examining preferences specified by a user of the client device, ining contents of the object, examining communication network characteristics, choosing a set of transcoding ıeters, forming a transcoded form of the object, sending the transcoded form to the client, and/or examining ork characteristics including estimating network bandwidth between the server and the proxy as well as between oxy and the client, and/or estimating delay between the servers and the proxy as well as delay between the proxy ıe client device, and/or providing a feedback to the user about a level of transcoding performed on the object, r the step of examining including determining the size of the object, and/or determining dimensions of the object, r calculating the compression ratio of the object. If the object is of type image, dimensions of the image object ətermined by area of the image in square pixels, and the compression ratio is determined by the bpp ratio of the object. The present invention allows both store-and-forward and streaming transcoding, thus allowing forming scoded form before the step of receiving the object from the servers is complete. This method can be applied to , GIF and other image types. Another aspect of this invention is that it allows sending out the transcoded form ə the step of forming a transcoded form of the object is complete.

] There are several other considerations that are important. The above examples for the concepts of the present ion are usual for image and video, etc. The wide use of the Internet has shown the value of JPEG and MPEG ressed image data. Audio coded data also needs to be decompressed, mixed with special sound effects, merged ther audio data, edited and processed in the real domain. Similar implementations are performed for other in-al, commercial, and military applications

] This invention may also be provided as a process, an article of manufacture, apparatus, system, architecture r a computer product. For example, it may be implemented as an article of manufacture comprising a computer ə medium having computer readable program code means embodied therein for causing a computer to perform ethods of the present invention.

] It is noted that although the description of the invention is made for particular arrangements of steps, the and concept of the present invention are suitable and applicable to other arrangements. For example, the in-ın is also adaptable to any browser although embodiment is directed towards web browsing only. Although primary deration is given to dynamic implementations, the invention may be employed with a combination of static, quasi-nic and dynamic implementations.

ıs

method for a transcoding proxy to facilitate browsing between a plurality of client devices and a plurality of servers ɔnnected via a communication network, the method comprising:

receiving a request from one of the client devices for an object stored at one of the servers,

forwarding the request for the object to said one of the servers,

receiving the object from said one of the servers,

examining preferences specified by a user of said one of the client devices,

examining contents of the object,

examining communication network characteristics,

choosing a set of transcoding parameters,

forming a transcoded form of the object, and

sending the transcoded form to said one of the clients.

2. A method as recited in claim 1, wherein the network characteristics includes bandwidth, and examining network characteristics includes estimating network bandwidth between said one of the servers and the proxy as well as between the proxy and said one of the clients.

3. A method as recited in claim 1 or claim 2, wherein the network characteristics includes delay, and examining network characteristics includes estimating delay between said one of the servers and the proxy as well as delay between the proxy and said one of the clients.

4. A method as recited in any preceding claim, further comprising providing feedback to the user about a level of transcoding performed on the object to form the transcoded form.

5. A method as recited in any preceding claim, wherein the step of forming a transcoded form employs dynamic adaptation.

6. A method as recited in any preceding claim, wherein the step of forming a transcoded form is started before the step of receiving the object from said one of the servers is complete.

7. A method as recited in any preceding claim, wherein the step of sending out the transcoded form is started after processing an initial fraction of the received object, and before the step of receiving the image object from said one of the servers is complete.

8. A method as recited in any preceding claim 1, wherein the step of sending out the transcoded form is started before the step of forming a transcoded form of the object is complete.

9. A transcoding proxy system for facilitating browsing between a plurality of clients and a plurality of servers connected via a communication network, the proxy comprising:

an HTTP proxy engine to receive a request from one of the clients for an object stored at one of the servers, and to fetch the object from said one of the servers,

an object transcoder to form a transcoded form of the object using a set of parameters for transcoding,

a dynamic policy module to determine the set of parameters of transcoding,

an image size and delay predictor module to gather characteristics of the object,

a user preference module to gather quality preferences specified by a user of said one of the clients, and

a bandwidth estimation module to estimate available network bandwidth,

wherein said dynamic policy module dynamically adjusts the parameters of transcoding using the input received from the image size and delay predictor module, user preferences module, and bandwidth estimation module for the purpose of improving satisfaction for the user, and the transcoding system provides feedback to the user about the level of transcoding performed.

10. A computer program comprising computer program instructions to cause a computer to perform the steps of the method as claimed in any of claims 1 to 8.

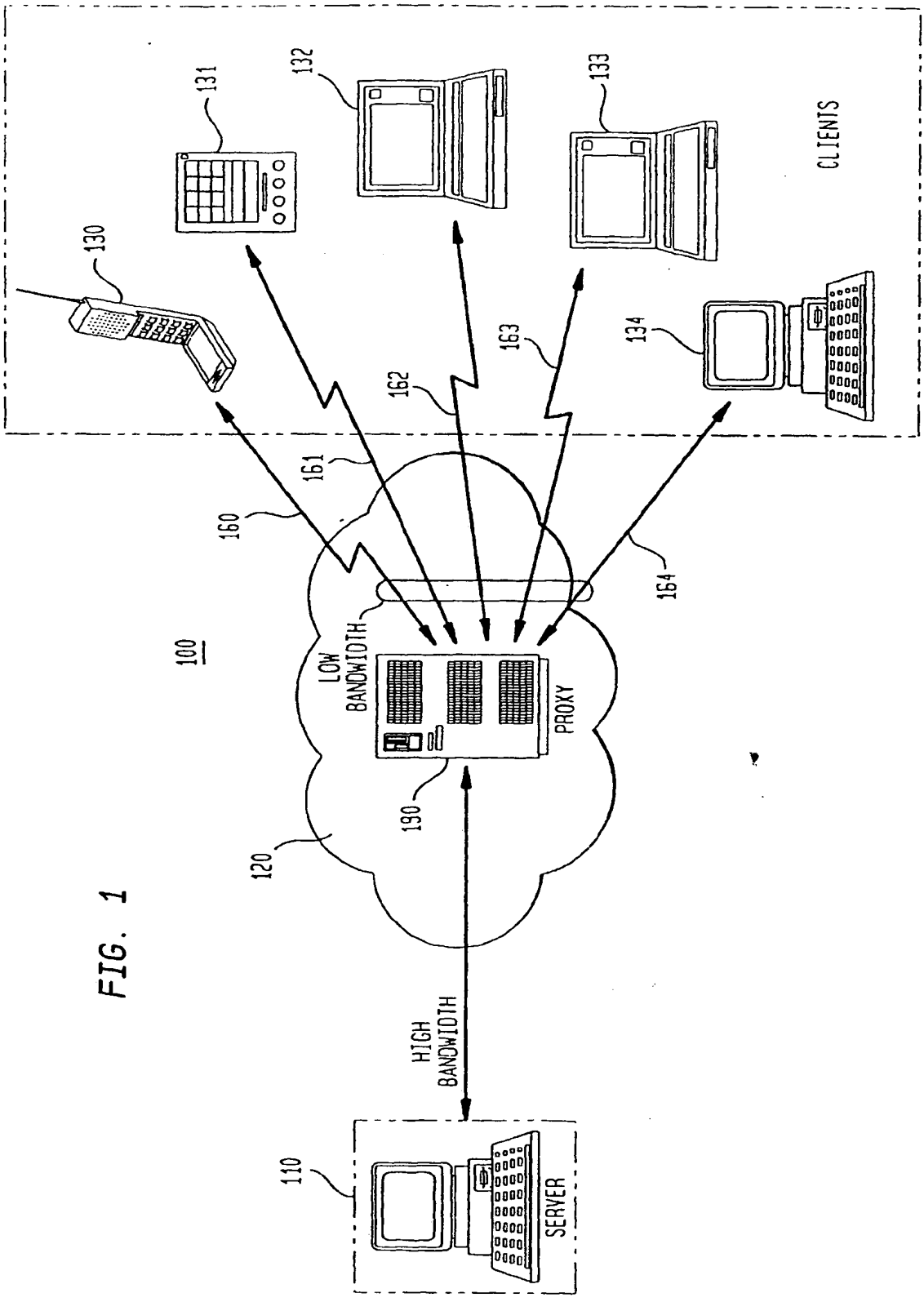


FIG. 1

FIG. 2

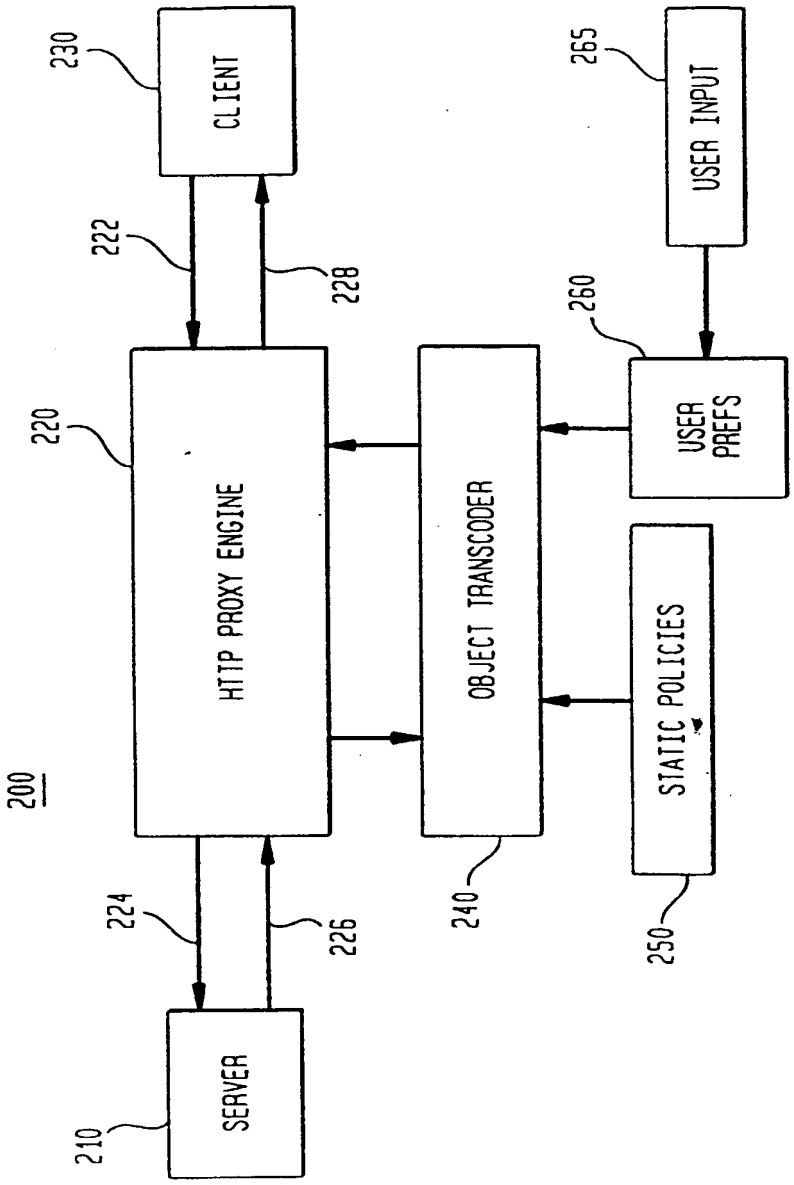


FIG. 3

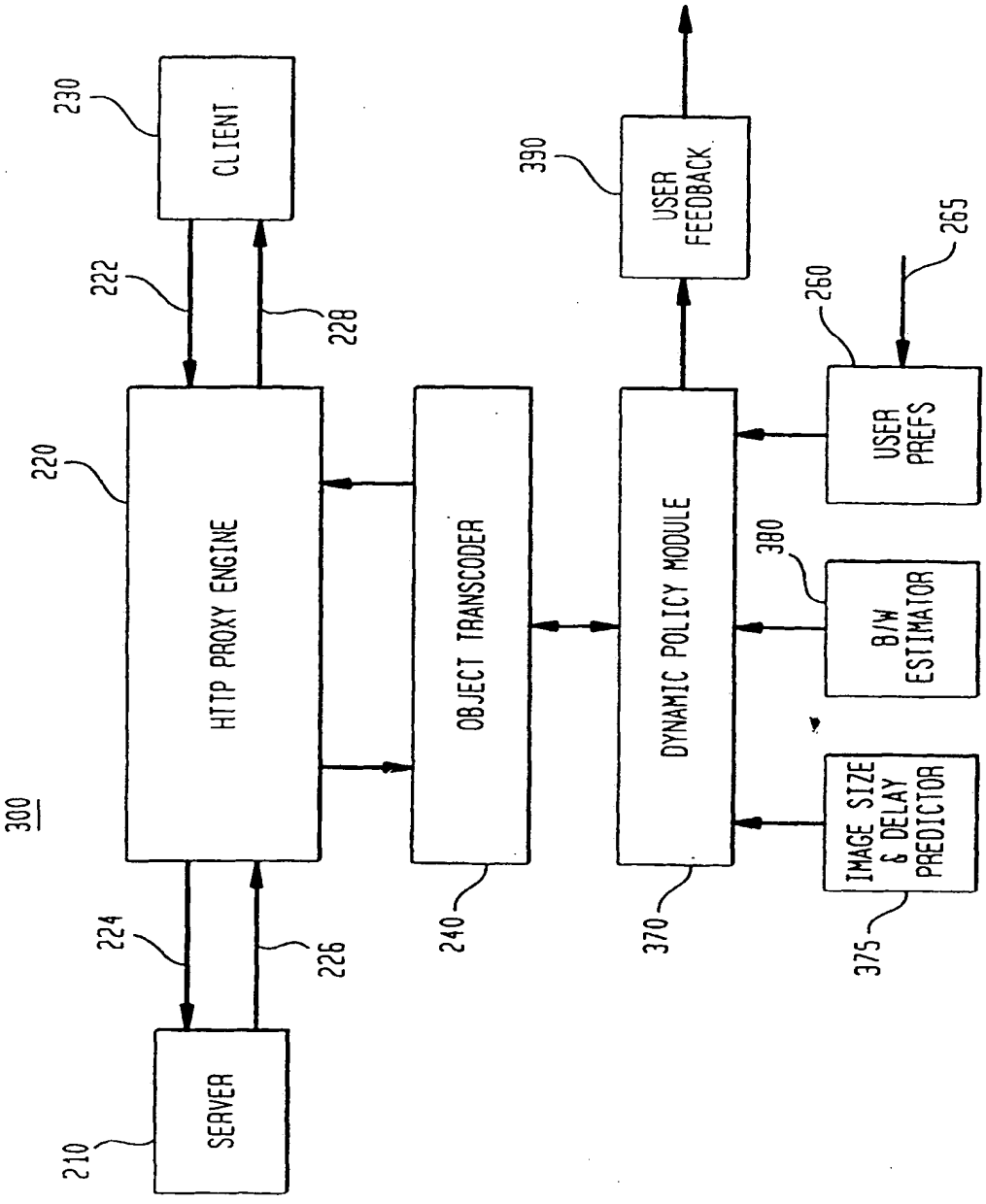


FIG. 4

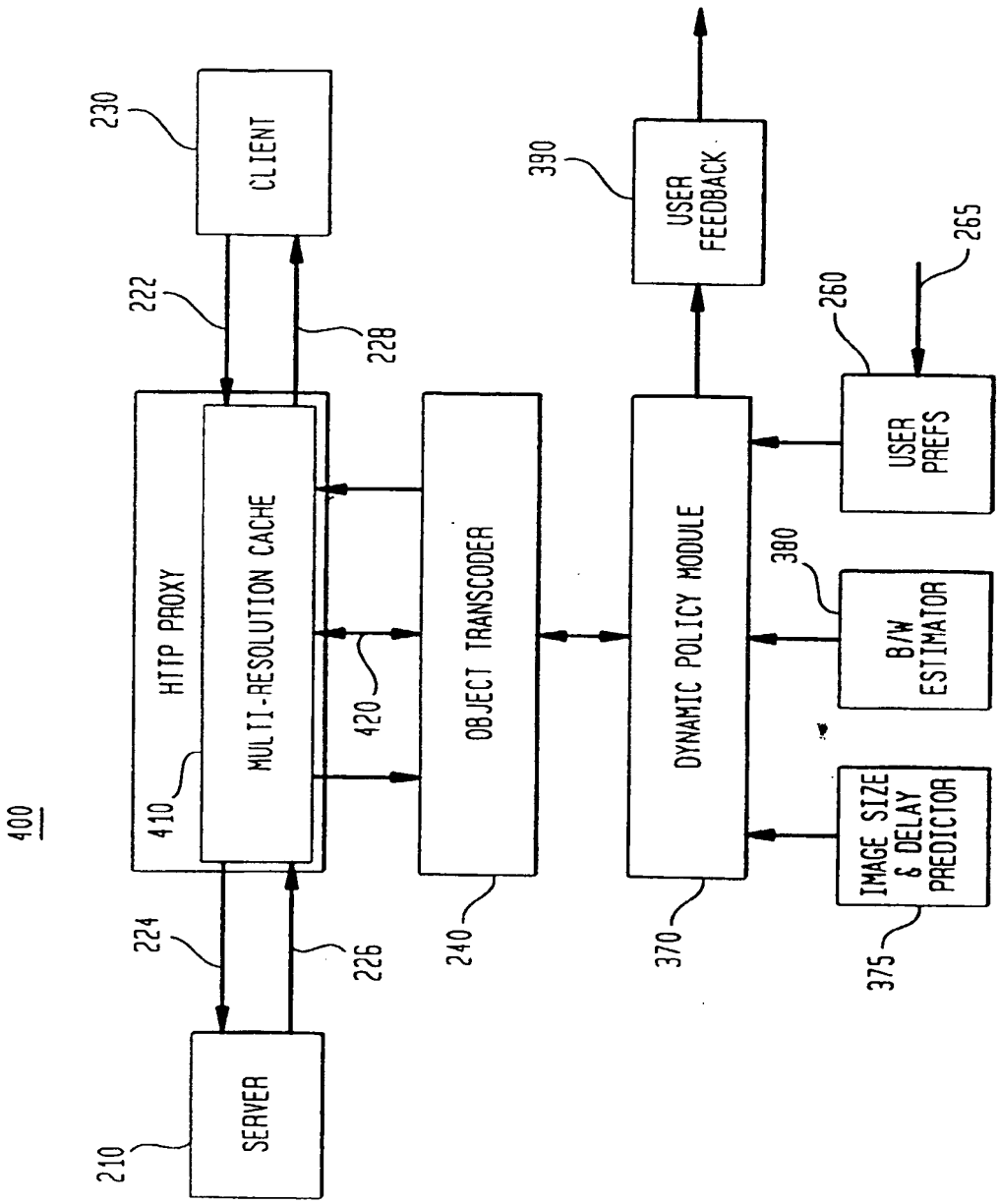


FIG. 5

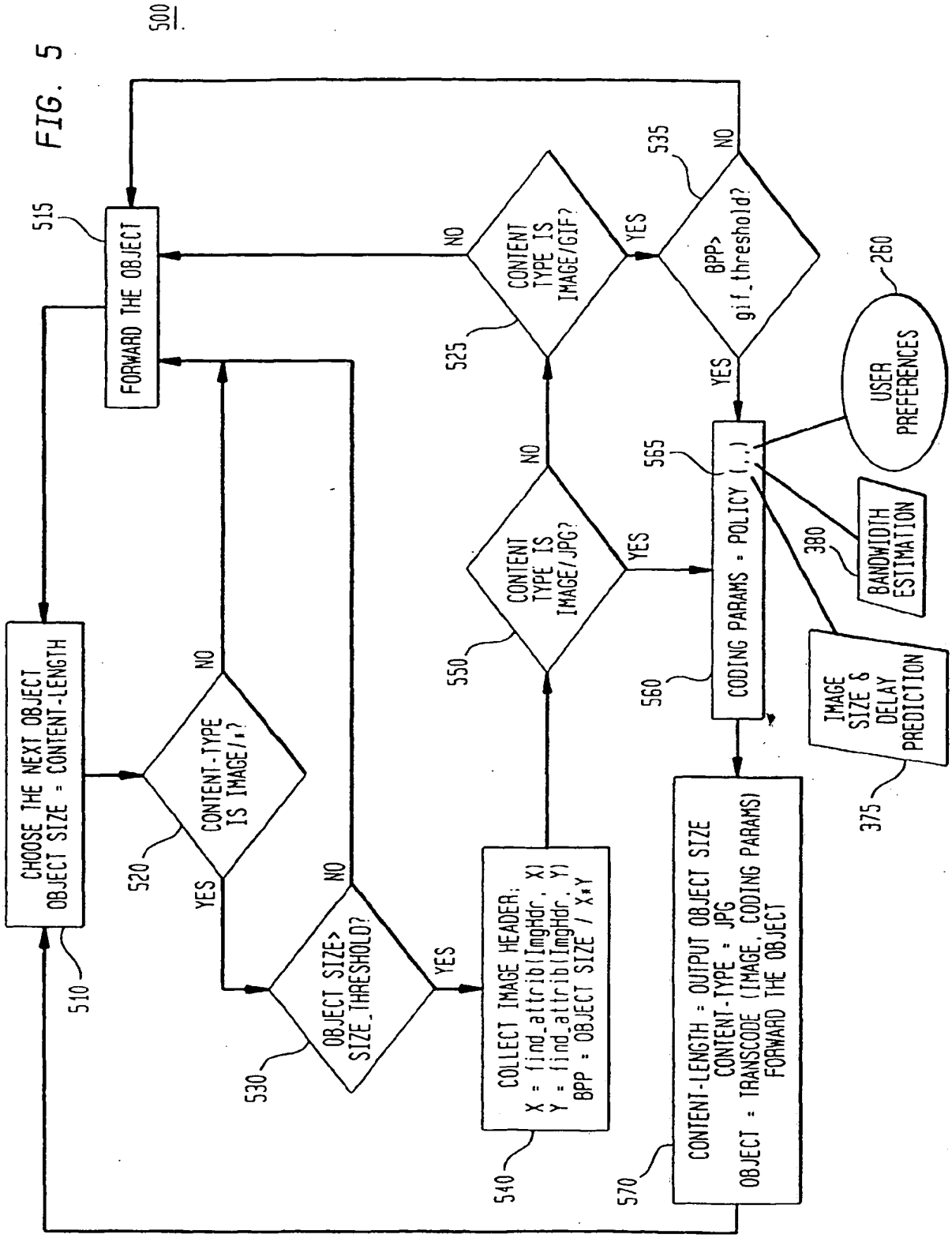


FIG. 6

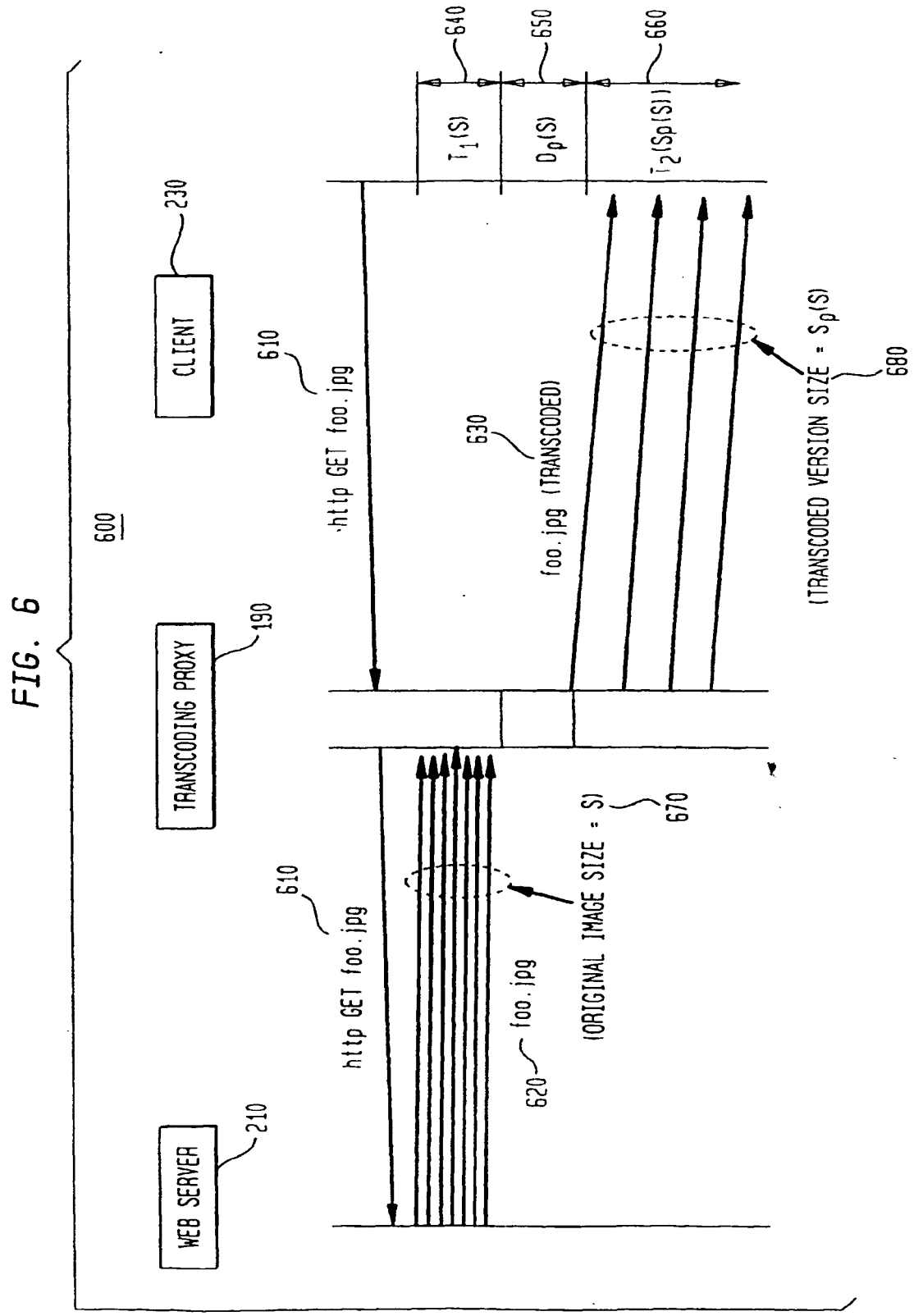


FIG. 7

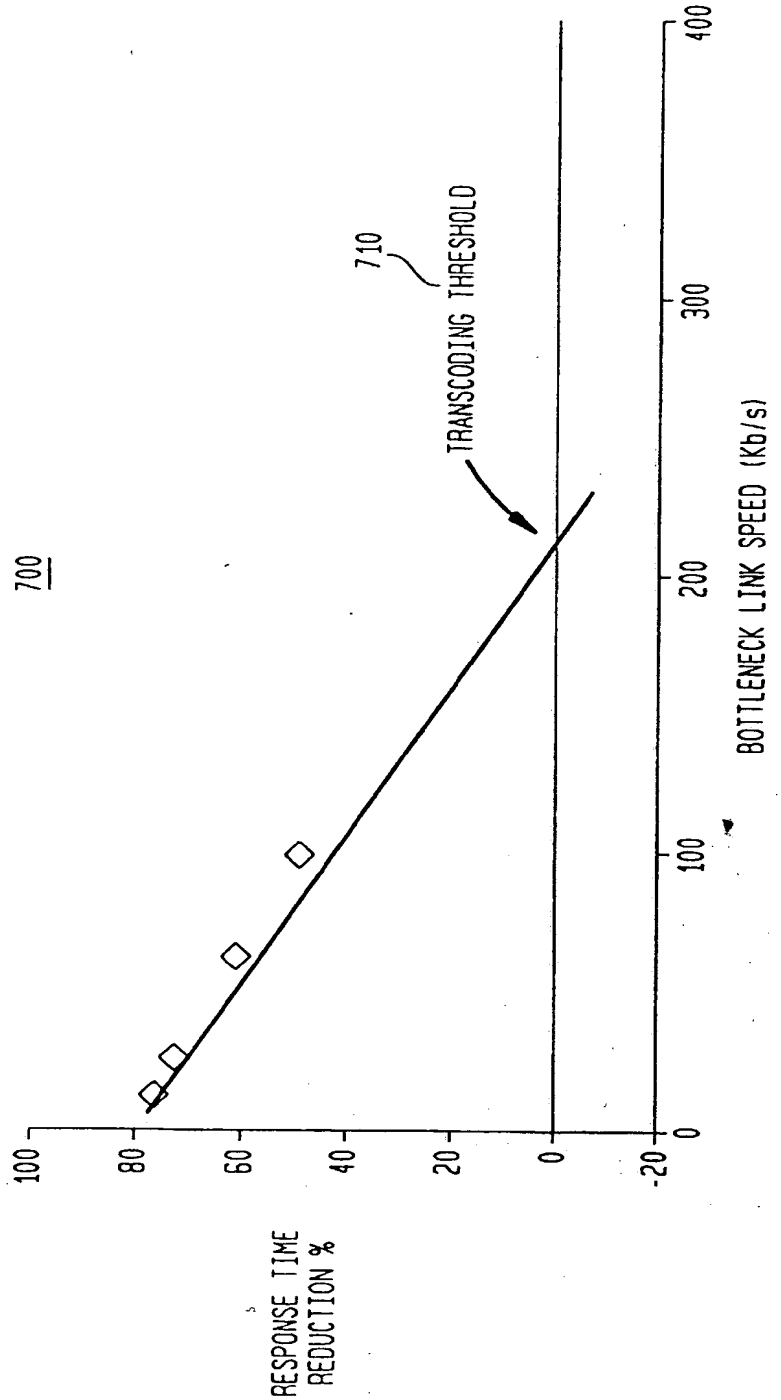


FIG. 8

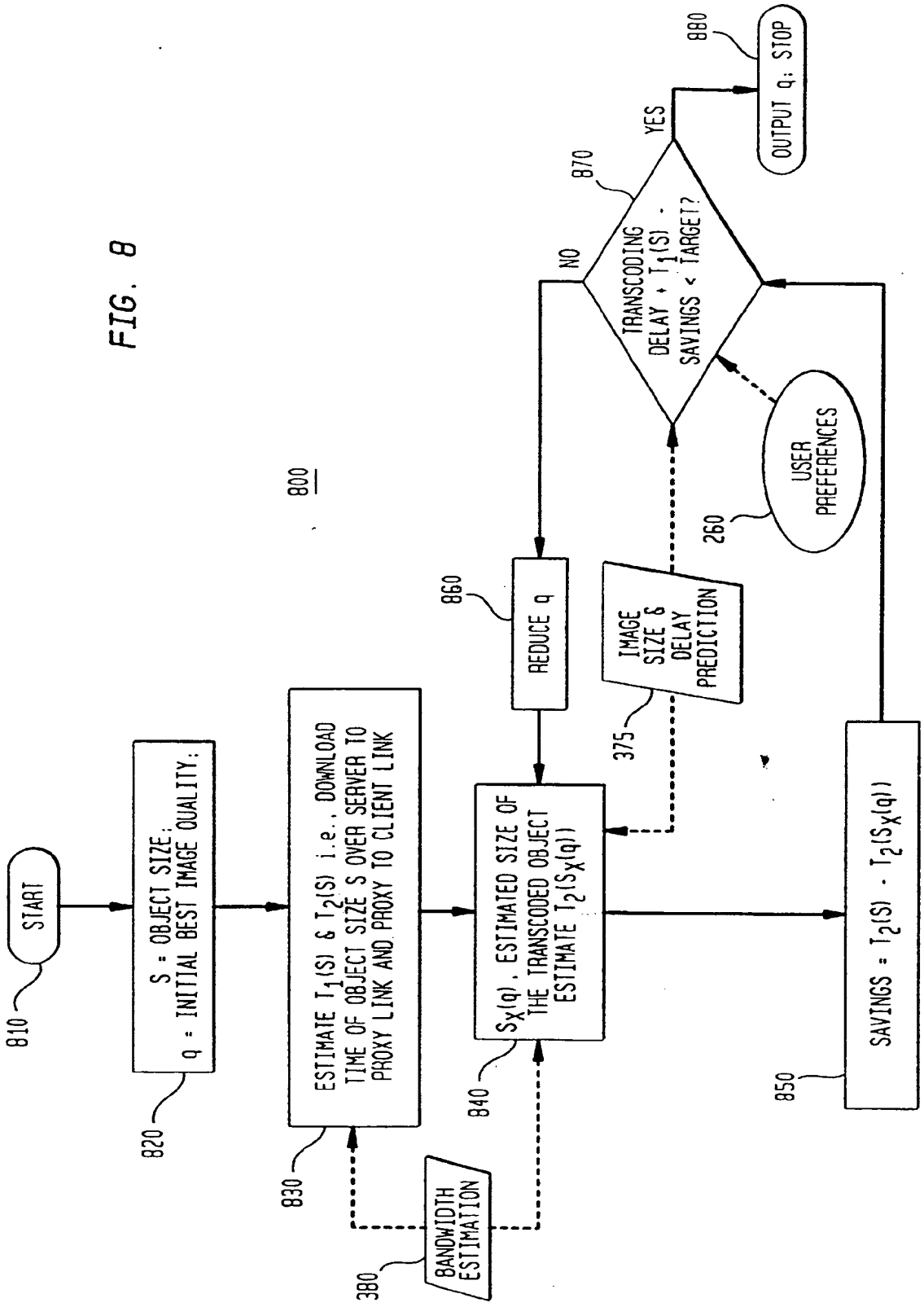


FIG. 9

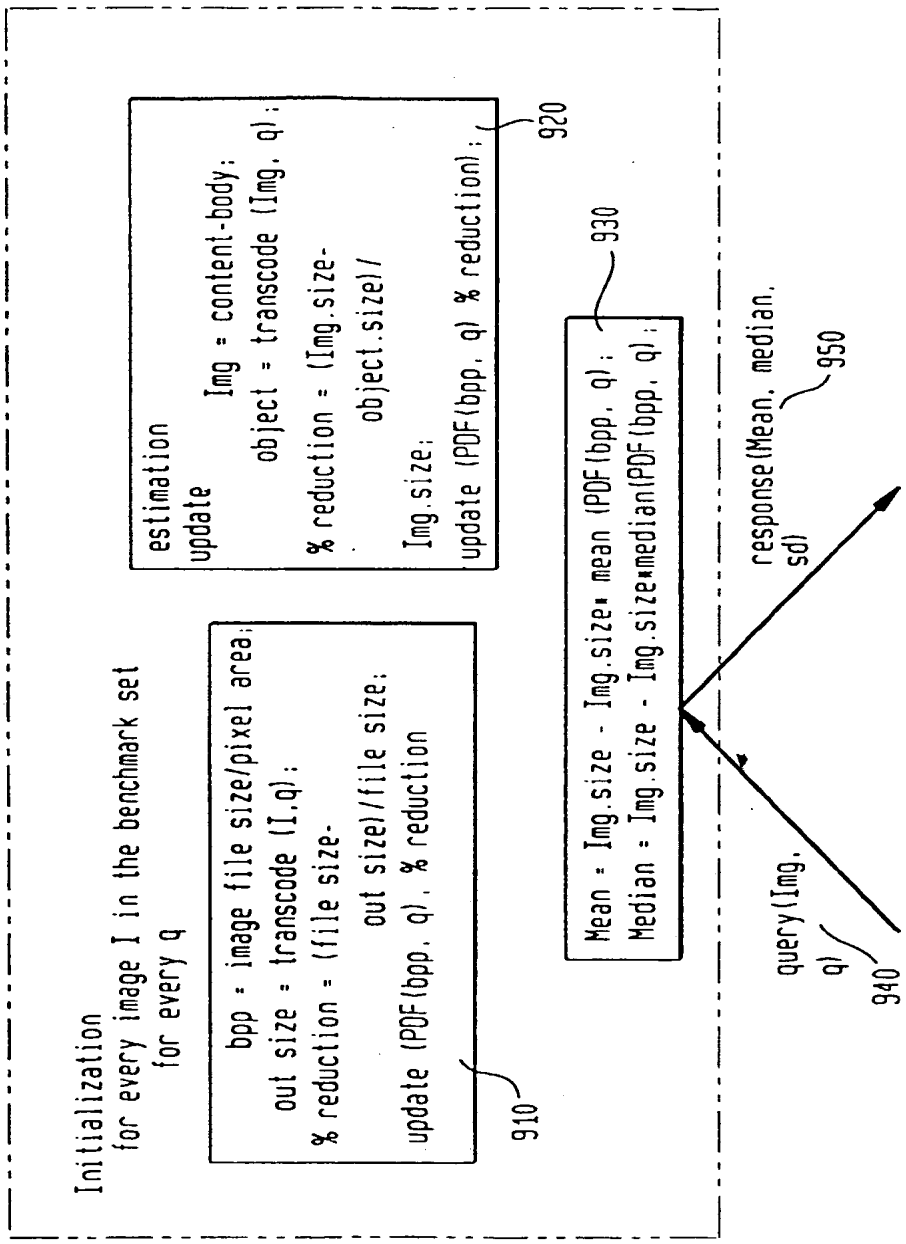


FIG. 10

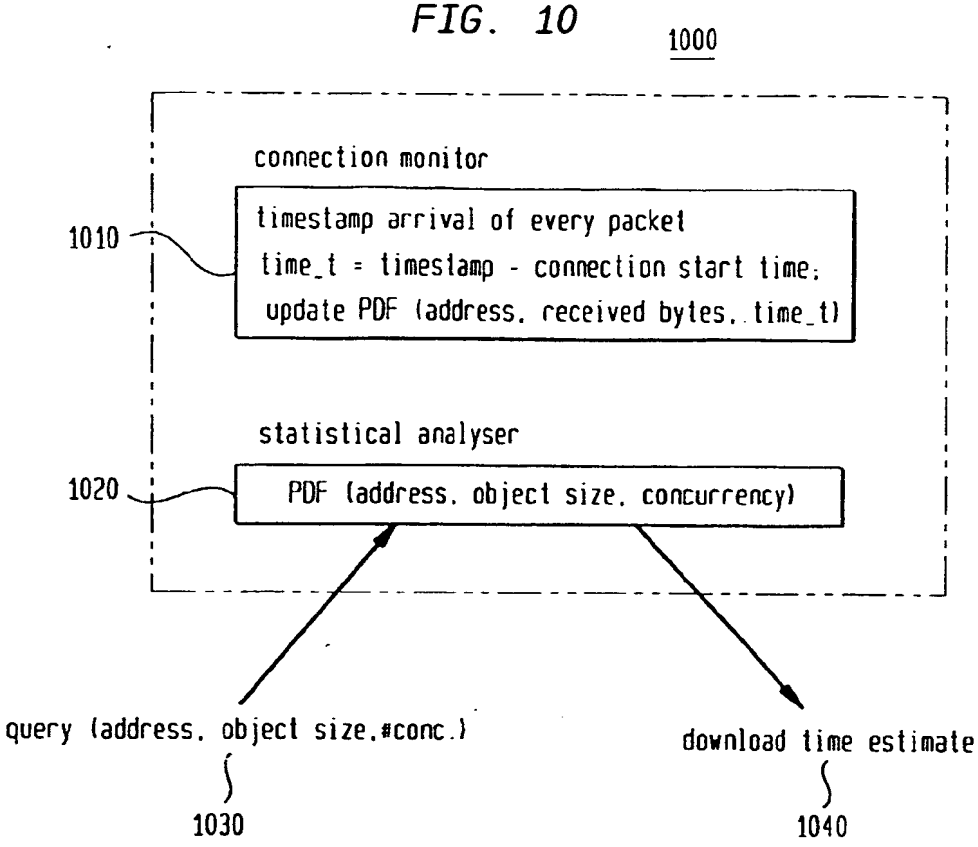


FIG. 11

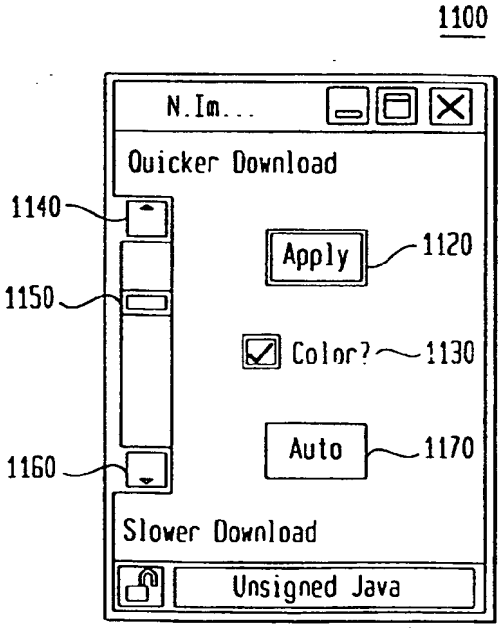
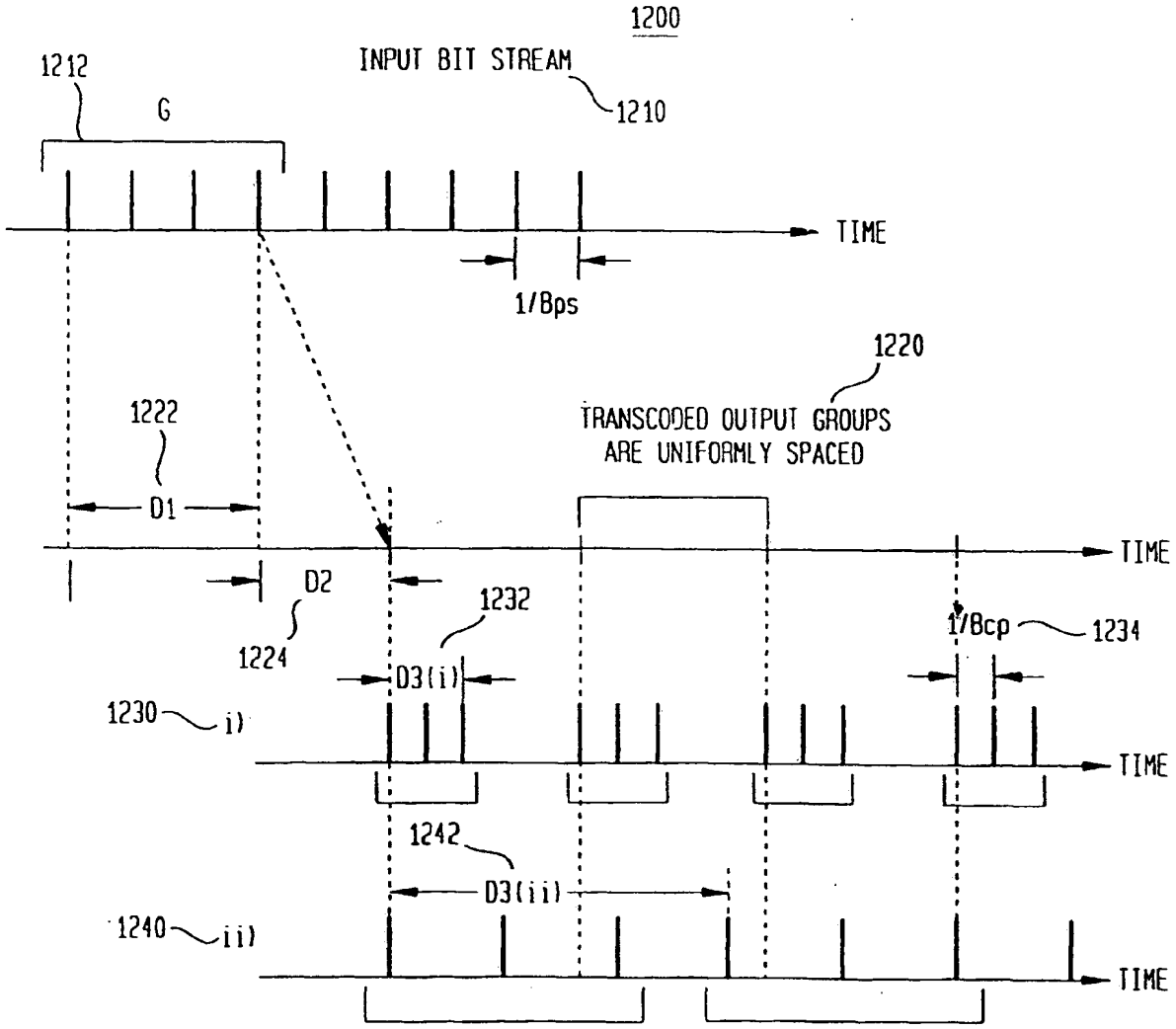


FIG. 12



(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 992 922 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
10.01.2001 Bulletin 2001/02

(43) Date of publication A2:
12.04.2000 Bulletin 2000/15

(21) Application number: **99307580.3**

(22) Date of filing: **24.09.1999**

(51) Int Cl.[7]: **G06F 17/30**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **29.01.1999 US 236622**
02.10.1998 US 102944 P

(71) Applicant: **International Business Machines Corporation**
Armonk, NY 10504 (US)

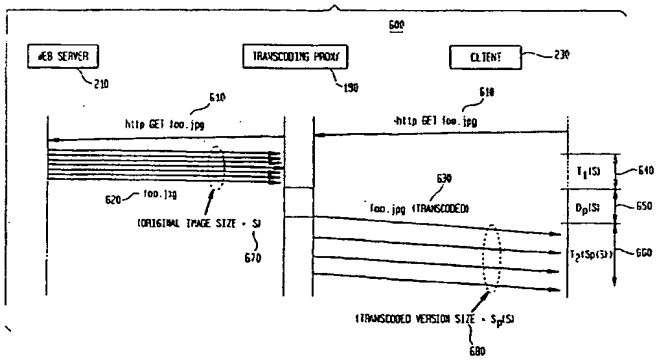
(72) Inventors:
- **Bhagwat, Pravin, c/o IBM United Kingdom Ltd.**
Winchester, Hampshire SO21 2JN (GB)
- **Han, Richard Yeh-whein,c/o IBM United Kingdom Ltd.**
Winchester, Hampshire SO21 2JN (GB)
- **La Maire, Richard O.,**
c/o IBM United Kingdom Ltd.
Winchester, Hampshire SO21 2JN (GB)
- **Mummert, Todd William,**
c/o IBM United Kingdom Ltd.
Winchester, Hampshire SO21 2JN (GB)
- **Rubas, James, c/o IBM United Kingdom Ltd.**
Winchester, Hampshire SO21 2JN (GB)

(74) Representative: **Davies, Simon Robert**
IBM,
United Kingdom Limited,
Intellectual Property Law,
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(54) Automatic image data quality adjustment to reduce response time of a Web server

(57) The present invention provides methods, devices and systems for dynamically adjusting transcoding parameters so as to increase the benefits of transcoding. Methods of adaptation are designed to cope with the variability of network characteristics and of the size of transcoded images. The invention also provides a method and apparatus to enable the transcoding proxy to adjust a quality-size tradeoff on a per-image and/or a per-client basis. The adaptive transcoder chooses different parameters for each object, and provides performance improvements. The invention further provides a general framework for making policy decisions taking into account available bandwidth, content and type of image, and user preferences. The invention also includes methods for generating feedback about the choice of optimal transcoding parameters to the user.

FIG. 6



EP 0 992 922 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 99 30 7580

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
P,X	HAN R ET AL: "DYNAMIC ADAPTATION IN AN IMAGE TRANSCODING PROXY FOR MOBILE WEB BROWSING" IEEE PERSONAL COMMUNICATIONS,US,IEEE COMMUNICATIONS SOCIETY, vol. 5, no. 6, 1 December 1998 (1998-12-01), pages 8-17, XP000790121 ISSN: 1070-9916 * abstract; figure 1 *	1-10	G06F17/30
X	WO 98 43177 A (INTEL CORP) 1 October 1998 (1998-10-01) * abstract; figures 1,3 * * page 3, line 6-14 * * page 4, line 5 - page 5, line 29 * * page 12, line 6 - page 14, line 22 *	1-10	
A	EP 0 811 939 A (WEBTV NETWORKS INC) 10 December 1997 (1997-12-10) * abstract * * column 1, line 13 - column 2, line 49 *	1-10	

TECHNICAL FIELDS SEARCHED (Int.Cl.7)

G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
MUNICH	14 November 2000	König, W

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 99 30 7580

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

14-11-2000

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9843177 A	01-10-1998	US 5902846 A	11-05-1999
		AU 6865698 A	20-10-1998
		BR 9811457 A	19-09-2000
		EP 1012733 A	28-06-2000
EP 0811939 A	10-12-1997	US 5918013 A	29-06-1999
		AU 3375197 A	05-01-1998
		JP 10228437 A	25-08-1998
		WO 9746943 A	11-12-1997
		US 6023268 A	08-02-2000
		US 5940074 A	17-08-1999
		US 6073168 A	06-06-2000
		US 5935207 A	10-08-1999
		US 5996022 A	30-11-1999
		US 5974461 A	26-10-1999

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82